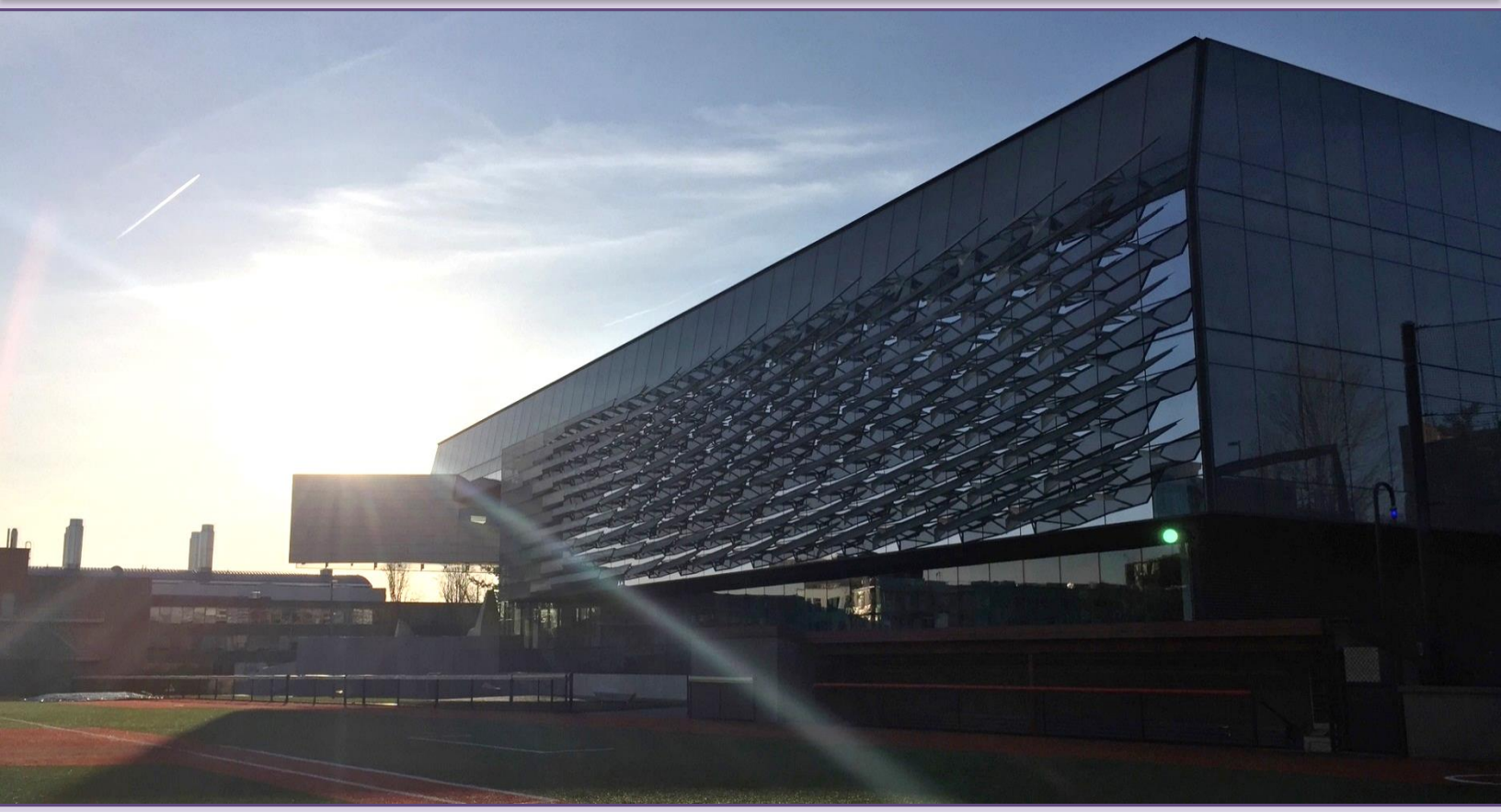
Cornell Data Science



Sponsorship Packet 2017

Contents

About us

Educate | Apply | Connect

We strive to **EDUCATE** people interested in data science at Cornell by providing learning opportunities. Our Introductory Data Science Training Program is an accredited course open to all Cornell students.

We want project team members to **APPLY** their knowledge meaningfully. We pride ourselves on having projects that are educational, competitive, and fun for our members.

We want to **CONNECT** members of the CDS community with other students, faculty, and industry leaders. We host talks given by distinguished Cornell faculty in CIS, hold intra-team socials, and organize company sponsored events.

Education

In Spring 2017, we began our **Introductory Data Science Training Program**.

This course, accredited by the College of Engineering, is a 11 week program designed to prepare any student, regardless of background, for a data analysis project. We equip them with statistical and analytical skills to work with data, which are used to extract meaningful results and accurate models. The course assumes no coding experience and teaches the R and Python languages.

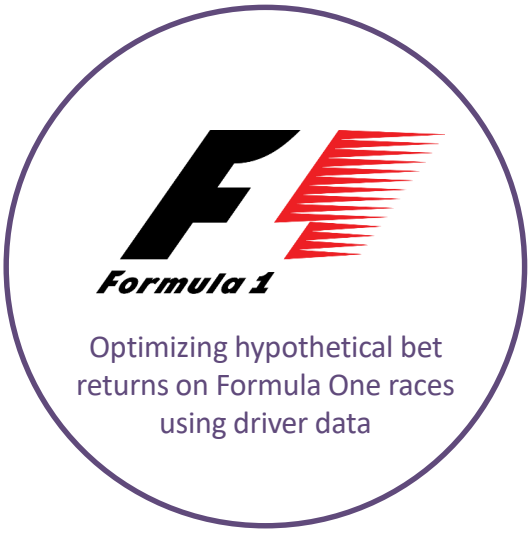
The education sub team designed a curriculum derived from our project team training series. With greater depth and focus on actual data projects as teaching material, the data science training program is proving to be one of the most effective ways for diving into data analysis here at Cornell. For Spring 2017, there were 120 students approved to take the course for credit.

Projects

Our **semester-based projects** are the core of CDS. Team members who join our projects benefit from the expertise of their project advisors (typically PhD students or professors) and fellow members, as well as organizational support from our executive board. Additionally, our team members have access to parallel computing and local cluster computing resources managed by frameworks such as Hadoop and Apache Spark.

Past Projects Include...

- NBA Win/Loss Predictions
- Reddit Recommendation Engine
- EEG Signal Classification with SVMs
- Formula One Team Performance
- Computational Finance
- Trauma Medical Data Analysis



Current Projects

KAGGLE COMPETITION

The Kaggle subteam **makes predictive models for competitions on Kaggle.com**, where companies and researchers post their data. Statisticians and data miners from all over the world compete to produce the best models. For Spring 2017, we participated in two competitions..

Through feature engineering and statistical techniques, each team member implements various ML algorithms on the dataset, including logistic regression, word2vec, and random forests. Team members will then improve their models using optimization techniques. Finally, we use ensemble methods to combine all algorithms into a single "super-algorithm" that we submit to Kaggle.

Current Projects

ALGORITHMIC TRADING

The Algorithmic Trading subteam is focused on **finding opportunities to capitalize using data driven approaches**. By using financial data, this subteam hopes to formulate accurate predictions for various market sectors. We hope to achieve these goals by using well known algorithmic trading approaches, and then supplementing them with Machine Learning techniques. We hope to use clustering algorithms, among other unsupervised learning algorithms, to find equities that we wish to trade. Additionally, we aim to use supervised learning techniques and Natural Language Processing techniques to create accurate predictions of future prices.

YELP DATASET CHALLENGE

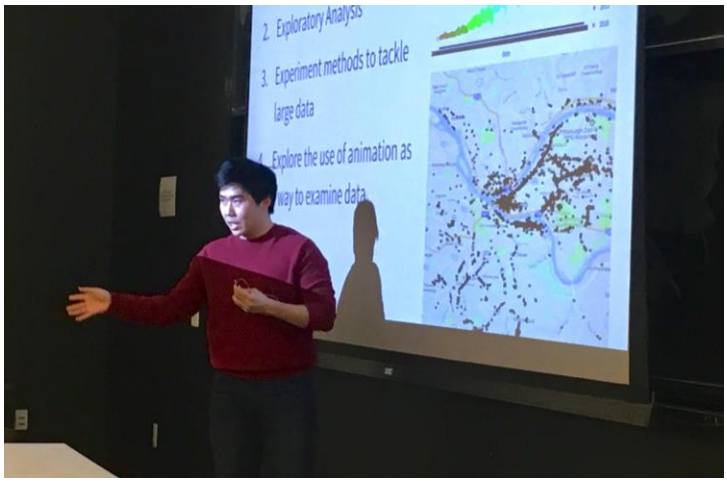
The Yelp subteam is a research team competing in the Yelp dataset challenge. Our goal is to **produce a rigorous research paper on bias factors affecting star-ratings in Yelp reviews**. The data we're working on includes detailed information on users, businesses, reviews, and data collected through Yelp features like check-ins and tips. The biggest challenge for the team is perhaps the sheer size and complexity of the data, which contains 15 gigabytes of JSON text. The size of the data requires a heavy emphasis on data engineering and parallel computing. Our team consists of students with strong machine learning and statistical backgrounds, and we are excited to explore areas like recommendation systems, time-series analysis, Natural Language Processing, and sentiment analysis.

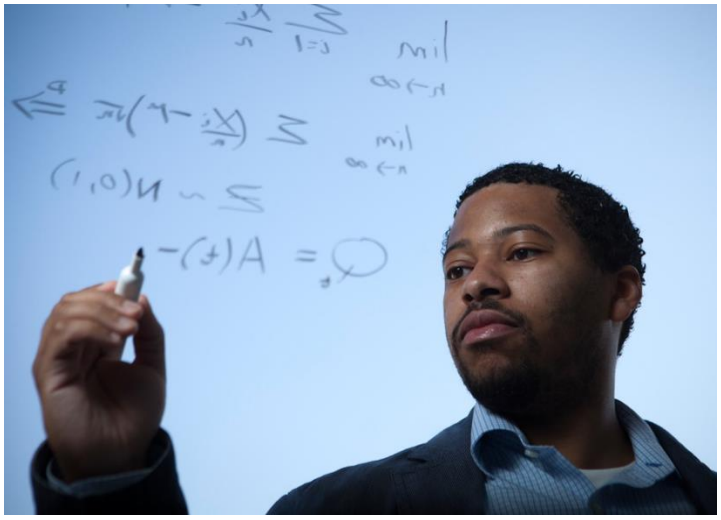
Events

Cornell Data Science hosts **events** with companies and professors to get our members acquainted with the world of data science. Past events have included:

Company events:

- **Microsoft tech talk** - A/B Testing: Putting the Science in Data Science
- **Yelp tech talk** - Modeling the competitive landscape: cost-per-click ad auctions
- **Ebay tech talk** - Ebay's Discovery Graph: providing personalized recommendations from sparse data





Professor Speaker Series:

- **Jamol Pender** - Stochastic analysis and optimal control of non-stationary queuing networks
- **Bobby Kleinberg** - Data Privacy: Challenges and Solutions
- **Kilian Weinberger** - Deep(er) Learning with Random Connections
- **Paul Ginsparg** - ArXiv and Adventures in Little Data

If your organization is interested in hosting an event with CDS, please email us at **cornelldatascience@gmail.com**.

Sponsorship

The members of Cornell Data Science are constantly seeking to expand and engage in new opportunities to achieve our 3 main goals, and we are always looking for new companies who can support our efforts. Below are our sponsor rewards and packages.

Sponsor Rewards:

Access to members

You will receive a resume book of all project team members. Out of over 150 applicants, a core group of 35 were chosen, having impressed our project managers with both their technical and team-oriented abilities.

Marketing

Depending on the selected package, company logos will be placed on all marketing materials, including:

- Facebook and Twitter pages
- Physical flyers, event banners, team shirts
- Slides before all on-campus events
- The official CDS website

Company Tech Talk

CDS will host your company for your own extensive tech talk. The presentation will be open to the public and likely well attended by the Cornell community. CDS will handle all marketing duties.

Sponsor Packages:

	Resume Book	Logo on Website	Tech Talk	Logo on shirt	Logo on all Marketing
Gold ($1000)	✓	✓	✓	✓	✓
Silver ($750)	✓	✓	✓		
Bronze ($500)	✓	✓			

Contact Us

Email: cornelldatascience@gmail.com

Mail: Cornell Data Science
B27 Upson Hall
Cornell University
Ithaca, NY 14853

Website: datascience.engineering.cornell.edu

Social Media:

Facebook.com/CornellDataScience

Linkedin.com/company/Cornell-data-science

Twitter.com/Cornell_Data

Medium.com/@Cornell_Data